３．５　国文学とコンピュータ

Personal e-Library System の思想

国文学研究資料館研究情報部教授
中　村　康　夫

はじめに

パソコンの性能も上がり、価格もずいぶん買いやすい設定になって、一人一人が持ちやすくなった。利用すればいろいろできることは、今や大方の感知しているところではある。しかし、実際にやってみようとすると、なかなかうまくいかない人がまだ多数おられるように見受けられる。

そうはいっても、国文学研究への利用について案内している本がないわけではない。例えば『電脳国文学―インターネットで広がる古典の世界―』（好文出版　2000 年 10 月）は、かなり多くの国文学研究者を視野に入れた好著といえる。しかし、これも手ごわく感じてしまう人はまだ多いように思われる。

「どういう知識を身に付ければ、どういうことができるのか」が具体的に想像できないと、人はなかなか本気になれない。知力の拡張は〈やる気〉を背景に必要とするから、〈ワザ〉を解説する前に、まず、結果をきっちり伝えることが必要なのだ。これを、ある程度のスケールの大きさを備えて、きっちり実行している〈案内〉本はまだない。つまり、こまごまとした〈ワザ〉の積み重ねの末に、どういうことが実現されなければならないかを、しっかり見据えている人物は、まだ、きわめて少数だということなのである。これは、言葉を換えれば、どういう研究がシステム化される必要があるか、また、可能かということが、まだまだ追求されていないということなのである。

ノートパソコンのハードディスクの容量として 20 ギガバイト程度が標準になりつつある今日は、作品本文をテキストで格納することを考えるならば、古典大系は言うに及ばず、新編国歌大観も、あれば群書類従も正続ともに入れてもまだまだ余裕があるわけで、できるのであれば、たいていの研究者はこの環境を自分のノートパソコンに実現しようとするのではないだろうか。

因みに、現在、国文学研究資料館から公開、提供されている旧日本古典文学大系のデータベースのサイズは、１冊約 30 万字（＝600KB）というから、100 冊として単純に 100 倍すると、60MB ということになる。1GB のハードディスクの容量があれば、1600 冊以上が記憶されることになる。これは、マーキング用のタグを含めてのサイズだから、テキスト本文だけならば、さらに１割増くらいになろうか。10GB なら、さらにその 10 倍になる。

『新編国歌大観』（全 10 巻）の本文データも、テキストでハードディスクに落とすと、ほぼ同じサイズ（約 70MB）であろうと思われる。

ここで、大事なことは、研究用に利用を限定する場合でも、著作権や利用約款は厳守されなければならないということである。

旧日本古典文学大系も『新編国歌大観』も、しかるべき手続き（利用申請と利用許可）や CD-ROM の購入という行為を果たして、限定された利用が可能になる。場合によっては、出版社やデータベース著作権者との直接の交渉が必要な場合もあるだろう。また、少なくとも、

いずれのデータベースも、そのままはもちろんのこと、加工したものでも、いかなる再配布もできない。つまり、加工しても、すべての権利が加工した側に移譲されるわけではないということである。

ノートパソコンが性能を拡大し、実行可能なことが拡大して、ハードディスク上にテキストデータとしては、小さな図書館サイズのデータが格納できることにはなった。しかし、著作権のことなどを考えると、フリーで流れているごくわずかなデータを除いて、流通させてはならないデータであったり、将来に渡って流通が期待できないデータであったりする。つまり、自力でデータ化（電子化）しなければどうしようもない情報が、必要なデータのかなりの部分を占めるということなのである。

必要なデータのうち、一部を電子化しないで、本の形で研究を進めることにすると、研究全体をシステム化することはできない。また、研究は、最新の成果を視野に入れていないといけないから、必ず、著作権等の関係から、流通していない情報を必要とする。結局、望まれることは、研究者個人が、個人の努力で、必要な情報を能率よく電子化し、データベース化して、次々の研究をシステム化していくためのシステム（Personal e-Library System）が必要になってくるのである。

まず、電子化情報を集めてこなければならない。

国文学研究資料館から公開、提供されている旧日本古典文学大系をダウンロードして、自分の手元で便利に検索、加工できるデータにして常備しておこうということを、手っ取り早く実現するには、旧日本古典文学大系のデータベースを使いやすいデータベースマスタの形にして、それを直接、国文学研究資料館から利用希望者に配布する（ダウンロードを可能にする）のが一番である。しかし、旧日本古典文学大系のデータは、今実現している提供の形に変更を加えても、いかなる再配布も流通も禁止している。これは、本の出版社である岩波書店との交渉の結果そうなっているので、そうである以上、ＩＤを獲得した利用資格者が、各自の努力で必要な変更を加えて、自分の研究目的にあったデータに育てなければならない。

フルテキストのデータは、grep で検索して利用しようとする人が多い。検索の速度は快適で悪くはない方法だが、grep で利用しようと考える場合は、grep で利用するのに適した形に加工するのが常識だと思う。とにかく何か見つかるからということで、ダウンロードしたままを grep で検索していこうとする人がかなり多くいるように感じられるのは、かなり問題だと私は思う。それに、grep で検索するなら、正規表現をきっちり勉強してほしい。

プログラムを書く勉強もそうだが、正規表現もむつかしくはないので、勉強したてはいい。しかし、どちらも、しばらく使わないと見事に忘れてしまう。これがしっかり身につくまでは、新しい言語を身に付けるのと同じ程度に反復、継続が必要なのである。何年も英語を勉強してきたのに一向に身につかない人があるように、そっけない情報として知識を記憶に残そうとするだけでは初めから限界が見えている。知識を生きた形で記憶に残すこととは、理解やその上に実現する感激などによって温度が与えられ、その温度とともに情報が格納されるからではないだろうか。プログラムや正規表現の勉強が温度を獲得しながら進むだろうか。もし、温度を獲得するとすれば、それは、〈本気〉〈やる気〉が伴ったときだろう。そして、それは、こういう研究がしたいと、はっきり目的が見えたときであるか、必要な基本情報だけでも常備しておこうと、my Library の必要性を認識したときだろうと思われるのである。

何が必要か

今日の最新の研究は、XML（eXtensible Markup Language）というルールでマークアップし、柔軟な定義を可能にした上で、多様な検索を実現しようというものである。少し前は、SGML（Standard Generalized Markup Language）というルールでマークアップしたデータに対して、Open Text というソフトで高速に検索するものであった。いずれも、フルテキストのデータベースを構築し、検索しようとする場合には将来性のある魅力的な方法に違いないが、残念ながら、情報処理の専門家がことに携わるという保証がないところでは、現実に稼動していない。また、そのシステムの価格は、個人のレベルで購入できるものではない。

では、Personal に利用できるシステムとはどういうものか。

ひとつには、タグ仕様を最低限に絞ることによって、システム自体をシンプルなものにする。タグを限定することは、言うまでもなく、多様性に限界を設けることになるが、データを一定のルールに基づいて作るには、簡便で、利用者に負担感が少ない。また、システムをシンプルにすることのメリットは、開発するのに費用がかからないということであり、購入するにも無理がない。

今すぐに入手できるシステムとしては、岩波書店から CD-ROM 出版されている国文学研究資料館古典コレクションシリーズ（『源氏物語（絵入）』『吾妻鏡』）に付いている散文検索システムである。このシステムにもバージョンがあるが、吾妻鏡データベースに提供されたバージョン 1.80 が一番新しい。

この検索システムは windows95、98、Me、NT、2000 のどれでも稼動するが、細かなところまで確認できているのは前３種である。

実現する電子図書館

１．旧古典大系のデータベースが登録されたら

いきなりだが、国文学研究資料館古典コレクションの散文検索システムをお目にかける。次の図１はこの検索システムに『古今著聞集』を登録して、ブラウザで開いたところである。今、「巻第一」の冒頭を出してみた。

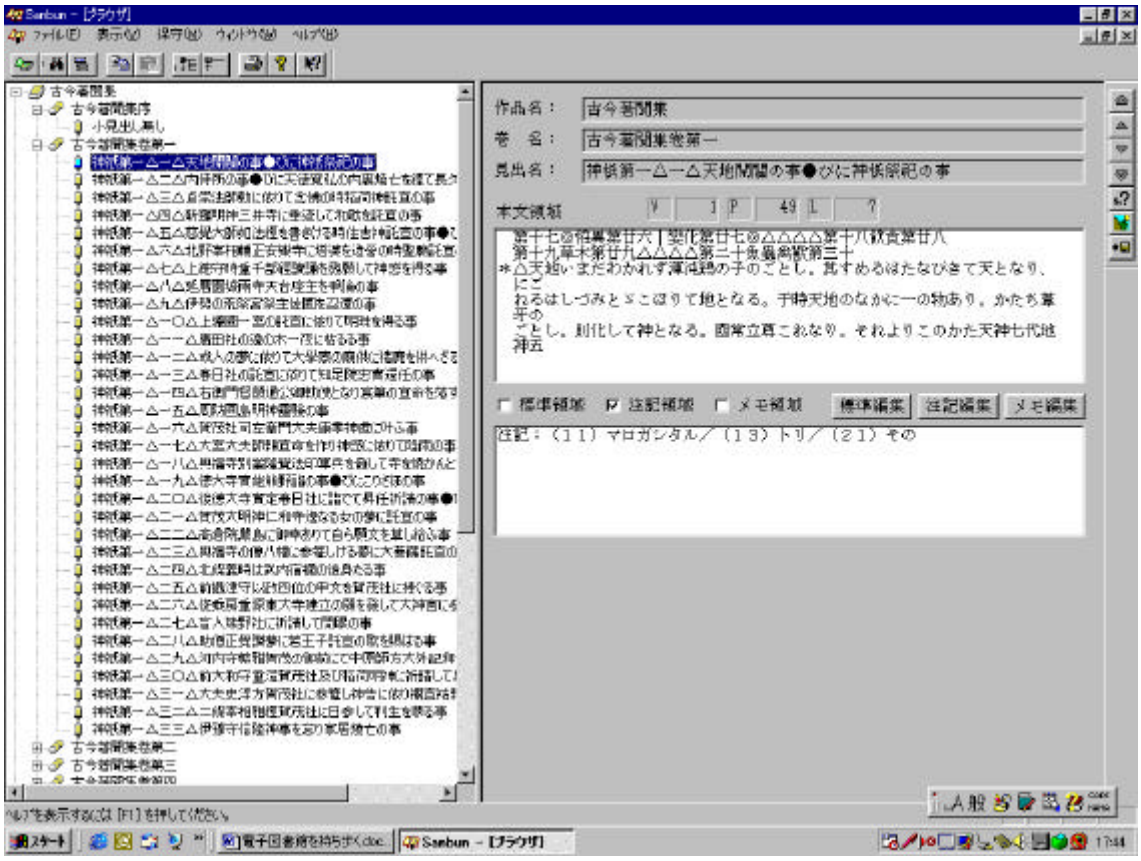

図１

ここにもっといろいろな説話作品を登録して、登録したすべてが左にリストで見えるようにすると、次の図２のようになる。

それぞれの作品名の下には巻が立っており、巻の下には各説話の見出しが立っている。全部の見出しを開けておいて、見出しを縦覧することも、一つの見出しが立った説話の本文を右に表示することもマウスの操作ひとつだ。そこで左を１回だけクリックするのか、ダブルクリックするのかは、ほぼwindowsの基本操作に準じている。感覚的に説明しておくと、１回目のクリックはそこを選択するという意味で、ダブルクリックの２回目のクリックはそこにジャンプするという意味になる。

今、登録した順に表示されているが、気になる人は登録する順番を考えて登録すればよい。検索するときには作品が選べるので、この程度の作品数ならば、とりあえずこのままでもいいかもしれない。

これはもともと『吾妻鏡』の検索システムではないかと不思議に思われるだろうか。このどこにも『吾妻鏡』は見えない。しかし、『吾妻鏡』は登録していないのかというとそんな

ことはない。ちゃんと登録されているし、いつでも表示したり検索したりできる。ではどこにあるのか。

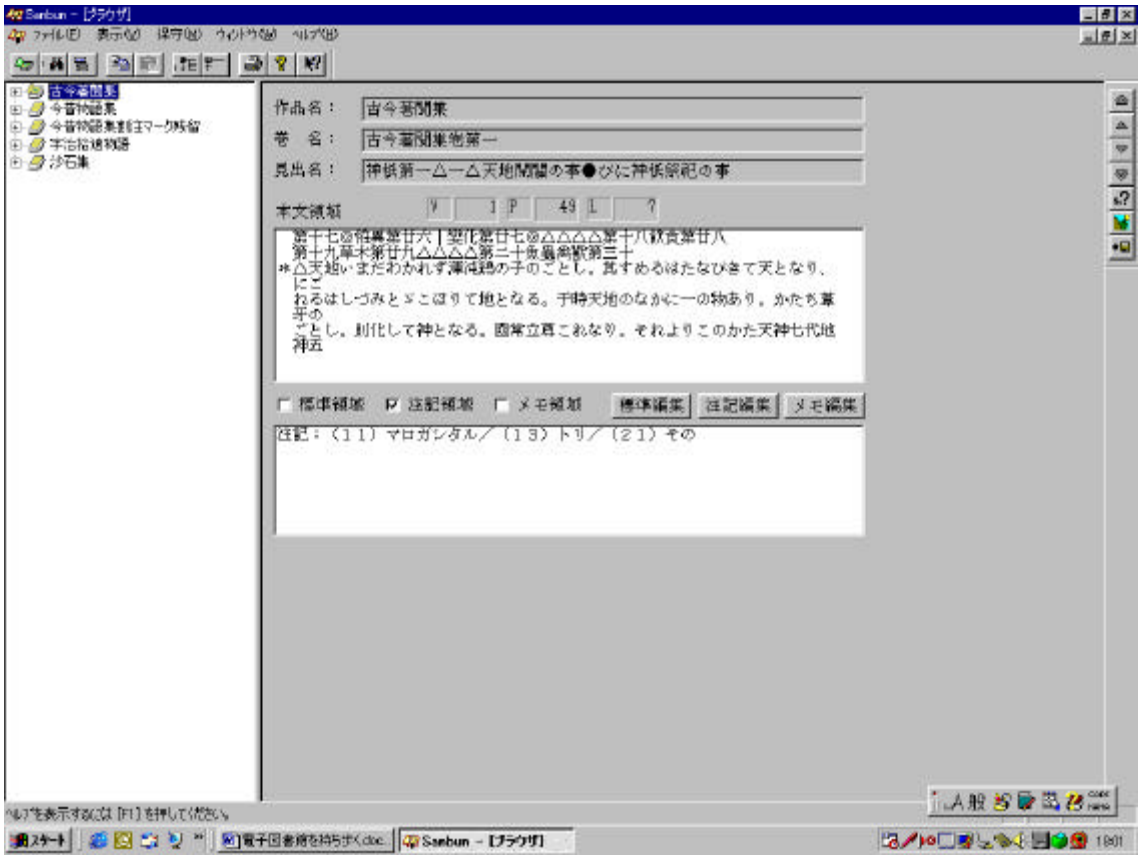

図２

ここで、少しだけ「フォルダ」について説明しなければならない。

この検索システムでは、データベースのデータはハードディスクに記憶する。そのときハードディスク上に一定の場所を占める。そしてその場所には場所の名前があって、その場所がフォルダであり、その名前がフォルダ名だ。

今ここでは、説話のデータベースのデータは「setsuwa」という名のフォルダに格納している。他にもジャンルによってさまざまなフォルダを作って、さまざまな作品のデータベースを登録して格納しておけばよい。

次の図３はデータベースのデータが格納されているフォルダが見えているところだ。

これはアルファベット順に表示されているので、『吾妻鏡』を登録した azuma というフォルダや、『保元物語』『平治物語』『平家物語』『太平記』などが登録されている gunki というフォルダなどはずっと上にある。ここでジャンルを選び、作品群を決めて、作品本文を縦覧したり、検索したりするという手順になる。

図３

図４

図３では「マスタフォルダの設定」となっているが、これは散文検索システムのメニューの中に「保守」という項目があって、その項目の中の一番上にある。それをマウスでクリックすると図４が見えるので、ここで一番上の「[..]」をダブルクリックするとマスタデータを格納してあるフォルダ（図３）がリストになって見えるというわけだ。

作ったデータベースマスタを、どういう名前のフォルダに登録するか、そのフォルダはどのタイミングで作るかなどの話は、本冊子のかなり後ろのほうでお話することになる。

古典コレクション散文検索システム

１．システムの基本

「散文検索システム」などといってしまうと、とにかくすでにあるものという印象を受ける。その印象に間違いはないが、そのように受身に考えるのではなく、何のためにそういうシステムがあるのかをよくお考えいただきたい。

それは、散文作品のフルテキストデータベースが完成できるとして、そのデータベースは、とりあえずどういうふうに使えればよいのかが大事なのだ。それ次第で、データベース設計も変わるし、システム設計も変わる。ここで大事なことは、データベースのプランが壮大すぎても実現が困難になり、システムを簡素にしすぎても役に立たないものになりかねないということだ。これならば、誰にとってもデータベースの構築は実現可能な程度で、この程度のシステム機能を実現しておくことが適当だと思われる、いわば、ほどほどのところを見極め、全体の構想にしなければいけない。そのような検討を経て、今日の姿を見るまでの経緯については、国文学研究資料館のホームページ（http://www.nijl.ac.jp/）から「組織」の「データベース室」を辿って、「懇談会報」のところをお読みいただきたい。

詳細は「懇談会報」に譲るとして、要するに、古典コレクション散文検索システムは、データベースを利用する人に共通して必要ということを考えて、作り出された基本システムなのだ。「とりあえず」などというのは、電子情報を使っての研究には未知の可能性を含めれば無限の多様性が想定されて、その個々の研究には更なる展開が予定されるからであって、個別の研究が必要とする手法までは、１つのシステムでは対応できないことが分かっているからだ。

散文検索システムは、大きく分けて次の２つの機能を持つ。

- データベースを登録する
- データベースを利用する

このうち、データベースを利用するためのシステム基本機能は、以下に示す４項目からなる。

- 利用するデータベースを選ぶ
- ブラウザで作品本文などを縦覧する。
- 作品本文などを検索する。
- 検索結果を利用する。

「作品本文など」というふうに「など」を付けるのは、作品の主本文以外の情報も、本文に関連付けて記述されていて、それらも同じように検索できたりするからだ。

「主本文以外の情報」というのは、例えば、漢字につけられた読み仮名や行間の注釈情報、さらには、掛詞の特性を１箇所に示すために複数の情報のうち主本文に示した以外の意味を傍記の形で示す場合、異表記や異文の指摘など、補助的あるいは補完的に本文の周辺に書かれている情報のことだ。図１右側下方に見えている「注記領域」のデータがこれにあたる。

４つの基本機能のうち、最初の「利用するデータベースを選ぶ」については、すでに前の章でふれたことになる。この章では、２つ目以降の機能について、図を中心に簡単に説明する。

２．ブラウザで作品本文などを縦覧する

ブラウザの機能は、メニューバーの左端にある「ファイル」から選ぶか、メニューバーの下に並ぶメニューボタンの左端にあるボタンをクリックする。

画面は左右の領域に分かれていて、左は分類項目の表示、右はデータの表示に使われる。

分類項目は、もっとも多いケースは

- 作品名
- 巻名
- 見出し

の３つに階層化されていて、『古今著聞集』の例でいえば、

- 古今著聞集
- 古今著聞集巻第一
- 神祇第一△一△天地開闢の事●びに神祇祭祀の事

がそれにあたる。これらの階層には上位・下位があって、並列ではないことがお分かりいただけるだろうか。

この階層化された分類情報は右のデータ表示領域にも再掲されていて、上部に３段で示されている。

分類項目は一様ではない。むしろ単純に３段に階層化して分類すると一般化して考えたほうがよい。きっちり階層化すると４段にも５段にもなる場合がある。例えば『吾妻鏡』の場合は、

- 作品名
- 年
- 月
- 日

のように４段が適当と考える方もおられるに違いない。しかし、このシステムでは３段しか許していないので、

- 作品名
- 年
- 月日

としている。むしろこの方が見た目にはしつこくなくてよいように感じられる。

右のデータ表示領域は、本文データに関していうと、位置情報とテキスト情報に分かれる。

位置情報は

V　　冊
P　　頁
L　　行

からなり、これは、旧古典大系の本に戻るための重要な情報にもなる。本に戻れれば、その箇所の頭注もすぐに開示できるし、そこから補注に回ることもできる。もちろん、頭注や補注はデータ化できるので、自分で書き足していってもよい。

テキスト情報は、次の４つの領域からなる。

本文領域作品の主本文を記述する
標準領域本文の表記を揃えてデータ化した領域
注記領域本文に関連する傍記等のデータ
メモ領域底本以外の文献情報やメモを記述する

本文領域は５行から１１行の幅で、自分の好みの大きさで表示できる。メニューバーの「表

示」から指定する。ただし、変更するときは、一端閉じて表示し直さないと新しい値にならない。

注記領域はマウスをクリックしてそこにチェックを入れると表示される。今回の方式では、その傍記がその行の何文字目の文字に対して書かれているかを「(n)」(nは整数)の形で示している。

標準領域とメモ領域は旧古典大系のデータには情報がない。もちろん、自分で情報を足していくのは自由だ。そのために編集ボタンが用意されているので活用されたい。

ブラウザで本文を表示した例は、図1、図2がそうなので、それをご覧いただきたい。

3．作品本文などを検索する

検索、検索結果の利用などについては、デモンストレーションで、ご説明する。